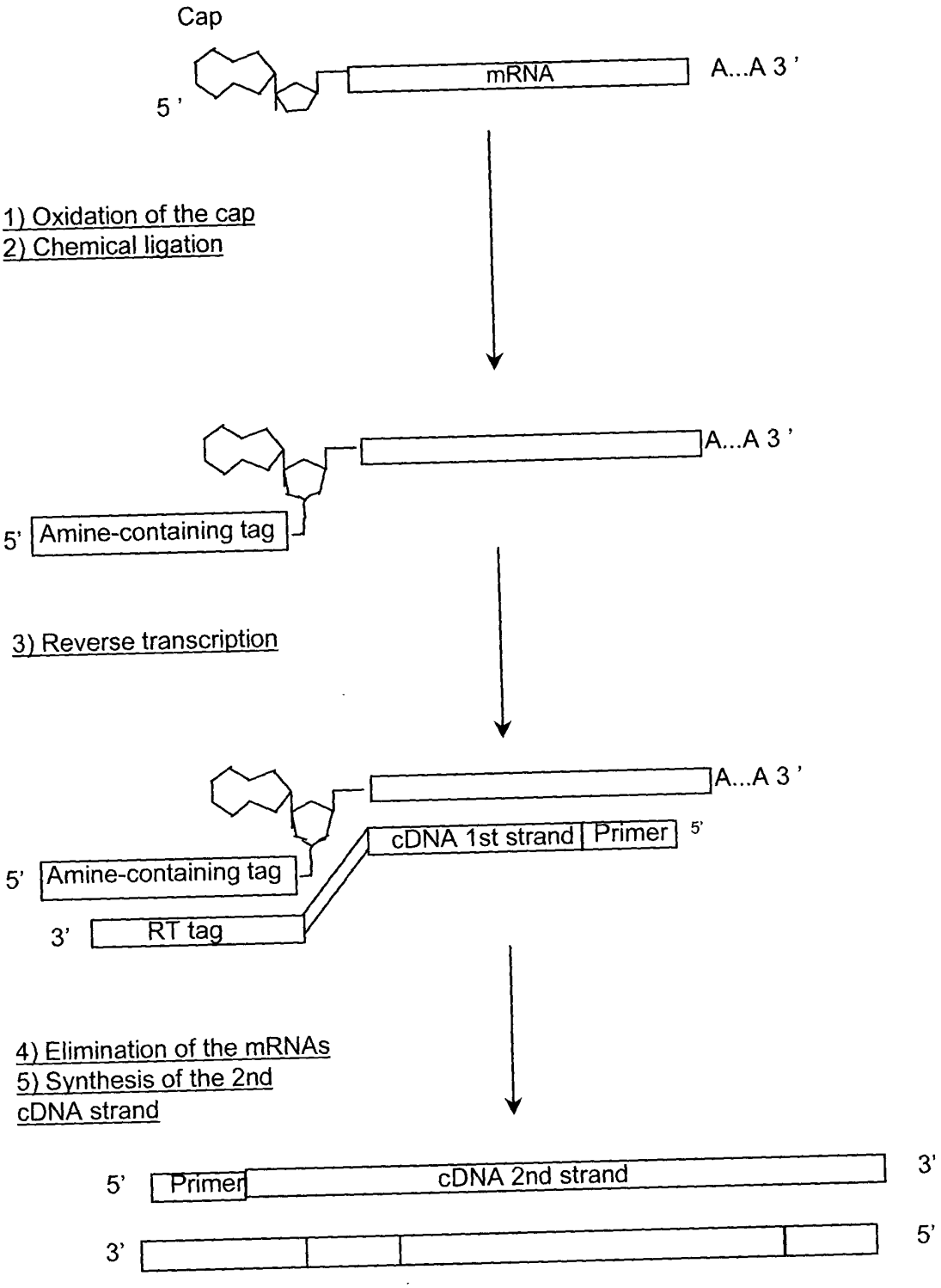

Cap
mRNA
A...A 3 '
5 '
1) Oxidation of the cap
2) Chemical ligation
A...A 3 '
5'
Amine-containing tag
3) Reverse transcription
A...A 3 '
cDNA 1st strand
Primer
5'
5'
Amine-containing tag
3'
RT tag
4) Elimination of the mRNAs
5) Synthesis of the 2nd
cDNA strand
5'
Primer
cDNA 2nd strand
3'
3'
5'

FIGURE 1

Minimum signal peptide score	false positive rate	false negative rate	proba(0.1)	proba(0.2)
3.5	0.121	0.036	0.467	0.664
4	0.096	0.06	0.519	0.708
4.5	0.078	0.079	0.565	0.745
5	0.062	0.098	0.615	0.782
5.5	0.05	0.127	0.659	0.813
6	0.04	0.163	0.694	0.836
6.5	0.033	0.202	0.725	0.855
7	0.025	0.248	0.763	0.878
7.5	0.021	0.304	0.78	0.889
8	0.015	0.368	0.816	0.909
8.5	0.012	0.418	0.836	0.92
9	0.009	0.512	0.856	0.93
9.5	0.007	0.581	0.863	0.934
10	0.006	0.679	0.835	0.919

FIGURE 2

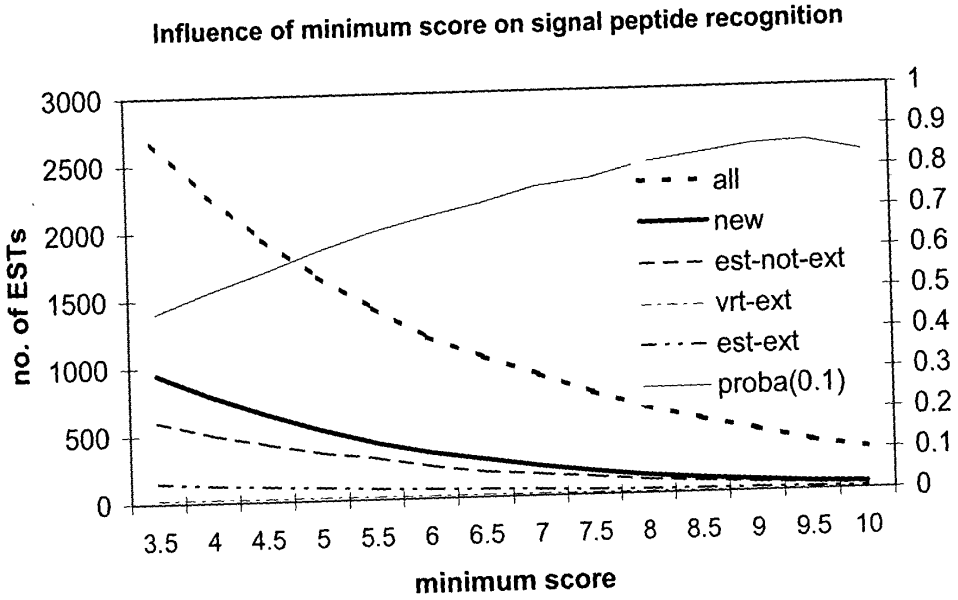

Influence of minimum score on signal peptide recognition
no. of ESTs
3000
2500
2000
1500
1000
500
0
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
all
new
est-not-ext
vrt-ext
est-ext
proba(0.1)
3.5 4 4.5 5 5.5 6 6.5 7 7.5 8 8.5 9 9.5 10
minimum score

FIGURE 3

Minimum signal peptide score	All ESTs	New ESTs	ESTs matching public EST closer than 40 bp from beginning	ESTs extending known mRNA more than 40 bp	ESTs extending public EST more than 40 bp
3.5	2674	947	599	23	150
4	2278	784	499	23	126
4.5	1943	647	425	22	112
5	1657	523	353	21	96
5.5	1417	419	307	19	80
6	1190	340	238	18	68
6.5	1035	280	186	18	60
7	893	219	161	15	48
7.5	753	173	132	12	36
8	636	133	101	11	29
8.5	543	104	83	8	26
9	456	81	63	6	24
9.5	364	57	48	6	18
10	303	47	35	6	15

FIGURE 4

Tissue	All ESTs	New ESTs	ESTs matching public EST closer than 40 bp from beginning	ESTs extending known mRNA more than 40 bp	ESTs extending public EST more than 40 bp
Brain	329	131	75	3	24
Cancerous prostate	134	40	37	1	6
Cerebellum	17	9	1	0	6
Colon	21	11	4	0	0
Dystrophic muscle	41	18	8	0	1
Fetal brain	70	37	16	0	1
Fetal kidney	227	116	46	1	19
Fetal liver	13	7	2	0	0
Heart	30	15	7	0	1
Hypertrophic prostat	86	23	22	2	2
Kidney	10	7	3	0	0
Large intestine	21	8	4	0	1
Liver	23	9	6	0	0
Lung	24	12	4	0	1
Lung (cells)	57	38	6	0	4
Lymph ganglia	163	60	23	2	12
Lymphocytes	23	6	4	0	2
Muscle	33	16	6	0	4
Normal prostate	181	61	45	7	11
Ovary	90	57	12	1	2
Pancreas	48	11	6	0	1
Placenta	24	5	1	0	0
Prostate	34	16	4	0	2
Spleen	56	28	10	0	1
Substantia nigra	108	47	27	1	6
Surrenals	15	3	3	1	0
Testis	131	68	25	1	8
Thyroid	17	8	2	0	2
Umbilical cord	55	17	12	1	3
Uterus	28	15	3	0	2
Non tissue-specific	568	48	177	2	28
Total	2677	947	601	23	150

FIGURE 5

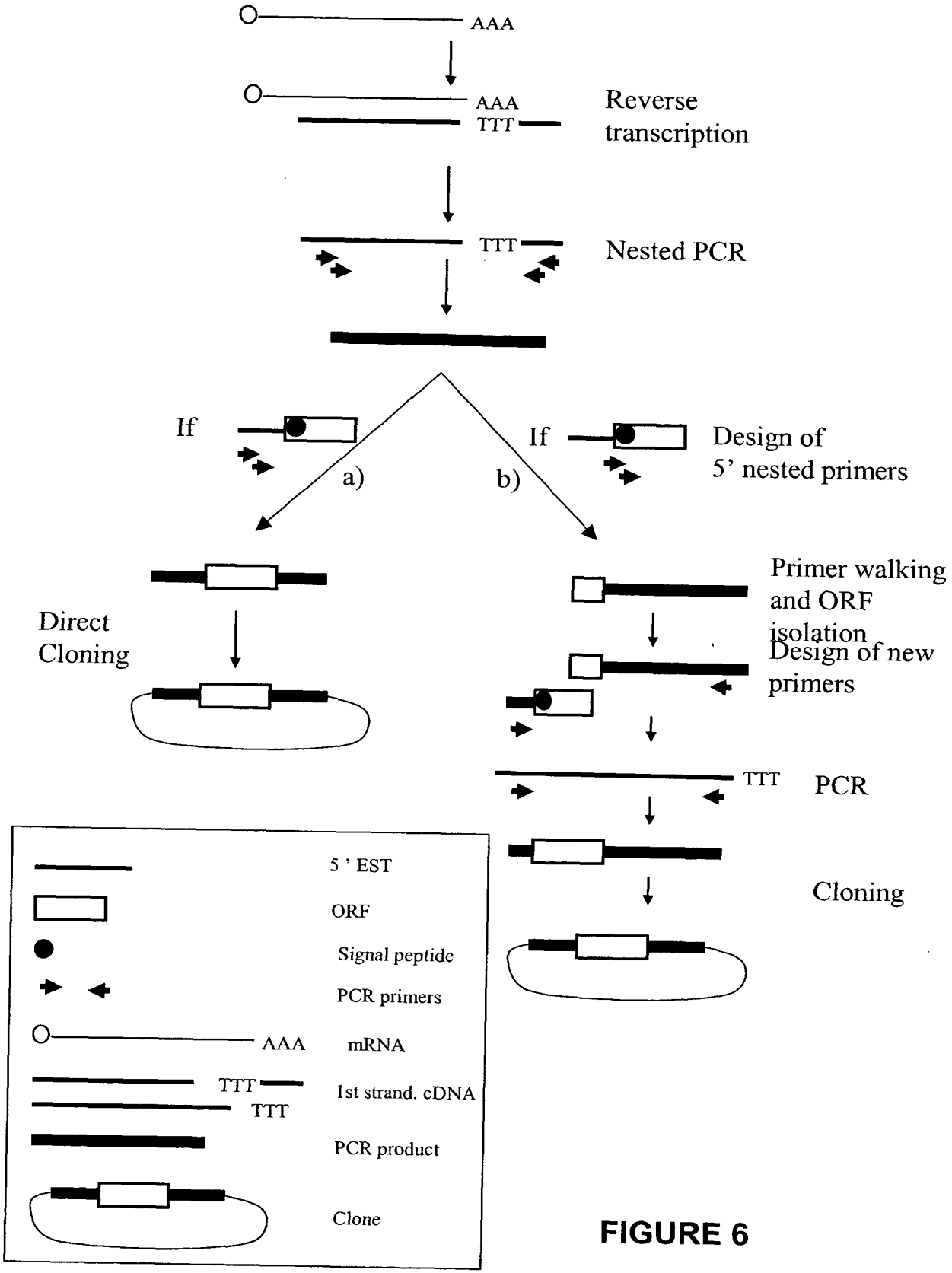
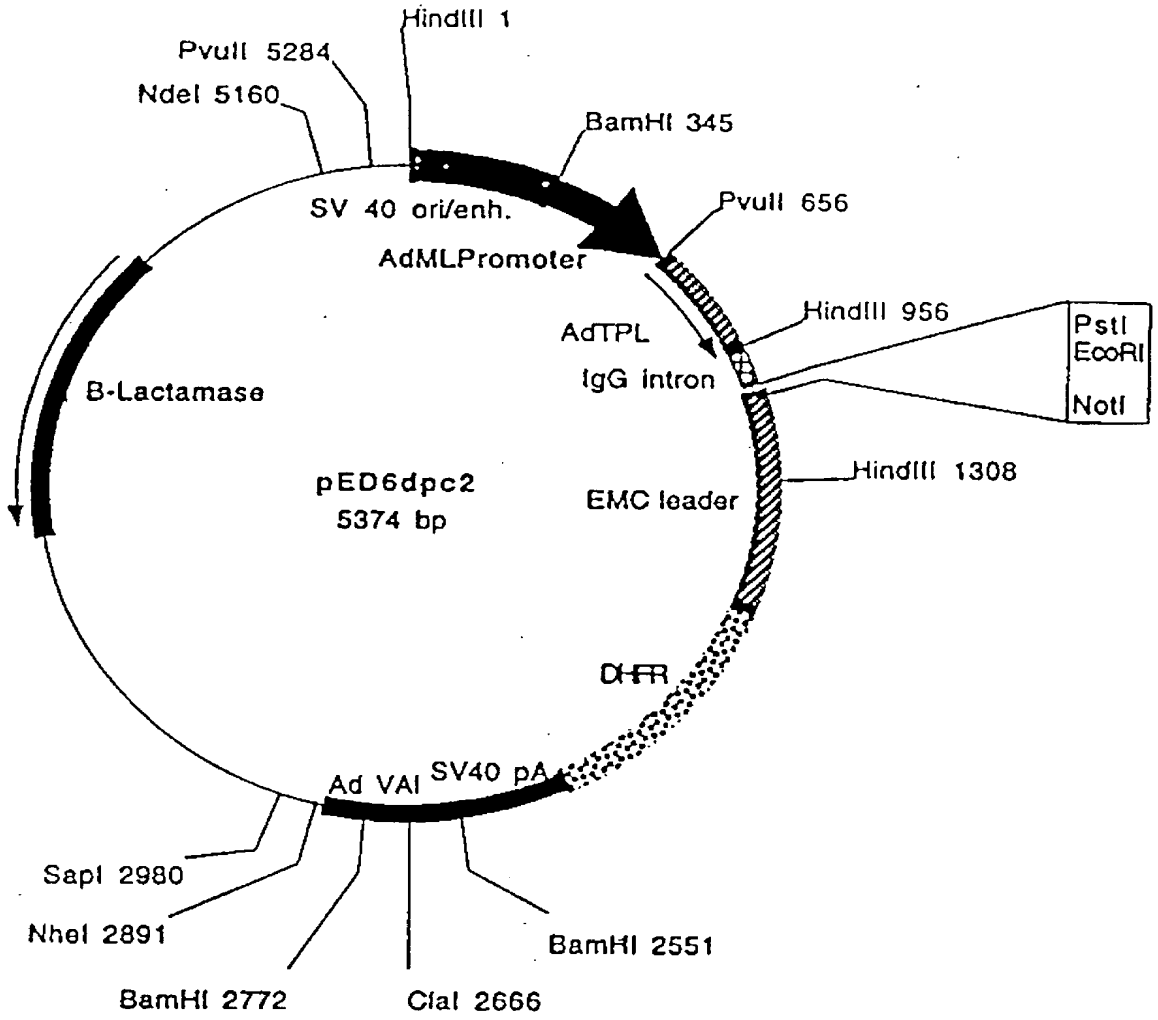

AAA
AAA
TTT
Reverse transcription
TTT
Nested PCR
If
a)
If
b)
Design of 5' nested primers
Direct Cloning
Primer walking and ORF isolation
Design of new primers
TTT
PCR
Cloning
5 ' EST
ORF
Signal peptide
PCR primers
AAA
mRNA
TTT
TTT
1st strand. cDNA
PCR product
Clone

FIGURE 6

7/14



Plasmid name: pED6dpc2

Plasmid size: 5347 bp

Comments/References: pED6dpc2 is derived from pED6dpc1 by insertion of a new polylinker to facilitate cDNA cloning. SST cDNAs are cloned between EcoRI and Not1. pED vectors are described in Kaufman et al. (1991), NAR 19:4485-4490.

FIGURE 7

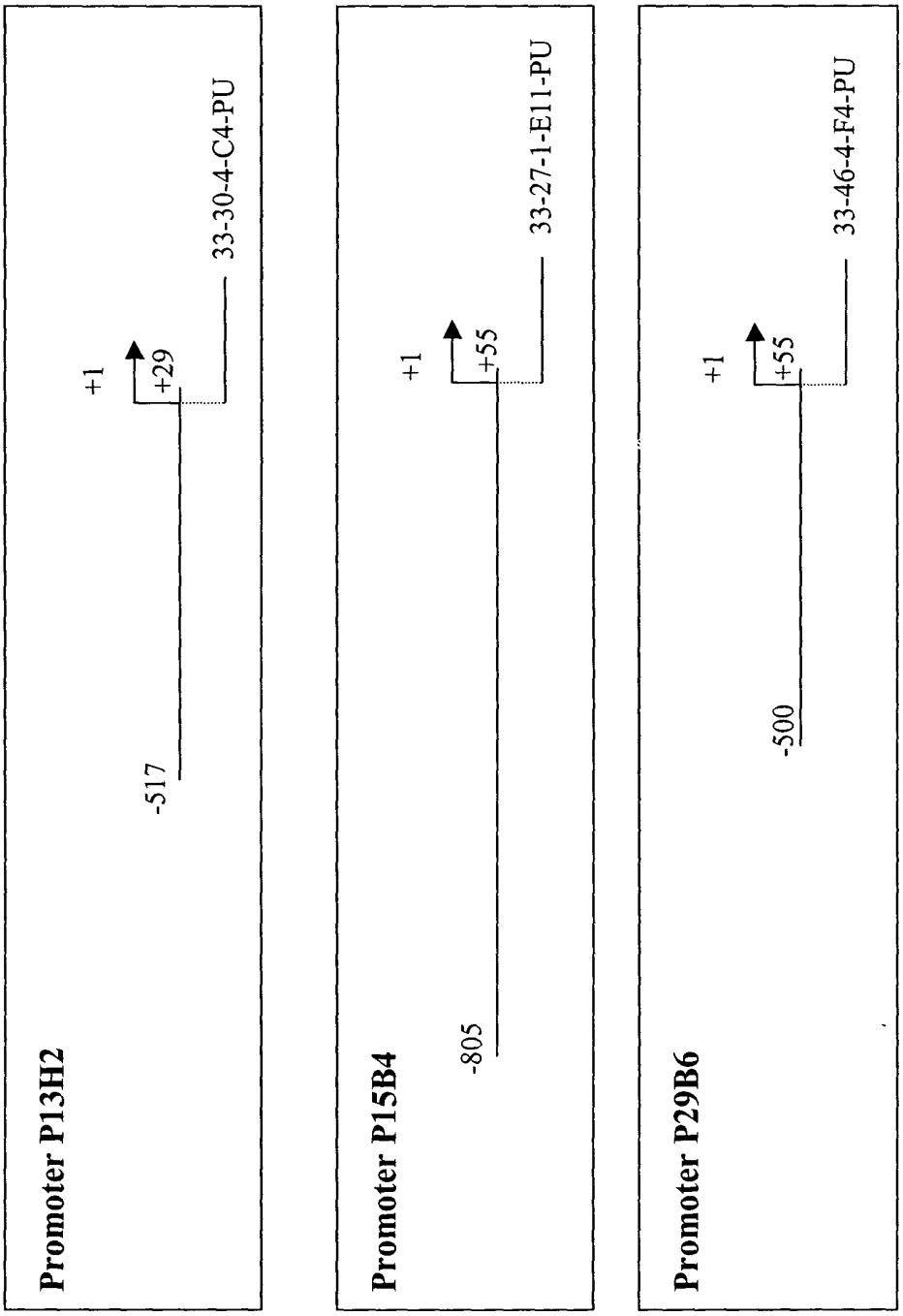


FIGURE 8

Description of Transcription Factor Binding Sites present on promoters isolated from SignalTag sequences

Promoter sequence P13H2 (546 bp):					
Matrix	Position	Orientation	Score	Length	Sequence
CMYB_01	-502	+	0.983	9	TGTCAGTTG
MYOD_Q6	-501	-	0.961	10	CCCAACTGAC
S8_01	-444	-	0.960	11	AATAGAATTAG
S8_01	-425	+	0.966	11	AACTAAATTAG
DELTAEF1_01	-390	-	0.960	11	GCACACCTCAG
GATA_C	-364	-	0.964	11	AGATAAATCCA
CMYB_01	-349	+	0.958	9	CTTCAGTTG
GATA1_02	-343	+	0.959	14	TTGTAGATAGGACA
GATA_C	-339	+	0.953	11	AGATAGGACAT
TAL1ALPHAE47_01	-235	+	0.973	16	CATAACAGATGGTAAG
TAL1BETAE47_01	-235	+	0.983	16	CATAACAGATGGTAAG
TAL1BETAITF2_01	-235	+	0.978	16	CATAACAGATGGTAAG
MYOD_Q6	-232	-	0.954	10	ACCATCTGTT
GATA1_04	-217	-	0.953	13	TCAAGATAAAGTA
IK1_01	-126	+	0.963	13	AGTTGGGAATTCC
IK2_01	-126	+	0.985	12	AGTTGGGAATTC
CREL_01	-123	+	0.962	10	TGGGAATTCC
GATA1_02	-96	+	0.950	14	TCAGTGATATGGCA
SRY_02	-41	-	0.951	12	TAAAACAAAACA
E2F_02	-33	+	0.957	8	TTTAGCGC
MZF1_01	-5	-	0.975	8	TGAGGGGA

FIGURE 9

Promoter sequence P15B4 (861bp) :					
Matrix	**Position**	**Orientation**	**Score**	**Length**	**Sequence**
NFY_Q6	-748	-	0.956	11	GGACCAATCAT
MZF1_01	-738	+	0.962	8	CCTGGGGA
CMYB_01	-684	+	0.994	9	TGACCGTTG
VMYB_02	-682	-	0.985	9	TCCAACGGT
STAT_01	-673	+	0.968	9	TTCCTGGAA
STAT_01	-673	-	0.951	9	TTCCAGGAA
MZF1_01	-556	-	0.956	8	TTGGGGGA
IK2_01	-451	+	0.965	12	GAATGGGATTTC
MZF1_01	-424	+	0.986	8	AGAGGGGA
SRY_02	-398	-	0.955	12	GAAAACAAAACA
MZF1_01	-216	+	0.960	8	GAAGGGGA
MYOD_Q6	-190	+	0.981	10	AGCATCTGCC
DELTAEF1_01	-176	+	0.958	11	TCCCACCTTCC
S8_01	5	-	0.992	11	GAGGCAATTAT
MZF1_01	16	-	0.986	8	AGAGGGGA

FIGURE 9 (cont)

Promoter sequence P29B6 (555 bp) :					
Matrix	**Position**	**Orientation**	**Score**	**Length**	**Sequence**
ARNT_01	-311	+	0.964	16	GGACTCACGTGCTGCT
NMYC_01	-309	+	0.965	12	ACTCACGTGCTG
USF_01	-309	+	0.985	12	ACTCACGTGCTG
USF_01	-309	-	0.985	12	CAGCACGTGAGT
NMYC_01	-309	-	0.956	12	CAGCACGTGAGT
MYCMAX_02	-309	-	0.972	12	CAGCACGTGAGT
USF_C	-307	+	0.997	8	TCACGTGC
USF_C	-307	-	0.991	8	GCACGTGA
MZF1_01	-292	-	0.968	8	CATGGGGA
ELK1_02	-105	+	0.963	14	CTCTCCGGAAGCCT
CETS1P54_01	-102	+	0.974	10	TCCGGAAGCC
AP1_Q4	-42	-	0.963	11	AGTGACTGAAC
AP1FJ_Q2	-42	-	0.961	11	AGTGACTGAAC
PADS_C	45	+	1.000	9	TGTGGTCTC

FIGURE 9 (cont)

97.8% identity in 92 aa overlap

```
                        10         20         30         40         50         60
SEQ ID NO:120 MASLGHILVFCVGLLTMAKAESPKEHDPFTYDYQSLQIGGLVIAGILFILGILIVLSRRC
              ::.: ::::::::::::::::::::::::::::::::::::::::::::::::::::::::
SEQ ID NO:180 MAPLHHILVFCVGLLTMAKAESPKEHDPFTYDYQSLQIGGLVIAGILFILGILIVLSRRC
                        10         20         30         40         50         60

                        70         80         90
SEQ ID NO:120 RCKFNQQQRTGEPDEEEGTFRSSIRRLSTRRR
              :::::::::::::::::::::::::::::::::
SEQ ID NO:180 RCKFNQQQRTGEPDEEEGTFRSSIRRLSTRRR
                        70         80         90
```

FIGURE 10

```
98.6% identity in 210 aa overlap

                                                        10        20        30
SEQ ID NO:121                                          MLTLLGLSLILAGLIVGGACIYKHFMPKST
                                                       :::::::.:::::::::::::.::::::
SEQ ID NO:181 LLSRTVRTQILTGKELRVATQEKEGSSGRCMLTLLGLSFILAGLIVGGACIYKYFMPKST
                    30        40        50        60        70        80

                         40        50        60        70        80        90
SEQ ID NO:121 IYRGEMCFFDSEDPANSLRGGEPNFLPVTEEADIREDDNIAIIDVPVPSFSDSDPAAIIH
              :::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::
SEQ ID NO:181 IYRGEMCFFDSEDPANSLRGGEPNFLPVTEEADIREDDNIAIIDVPVPSFSDSDPAAIIH
                    90        100       110       120       130       140

                       100       110       120       130       140       150
SEQ ID NO:121 DFEKGMTAYLDLLLGNCYLMPLNTSIVMPPENLVELFGKLASGRYLPQTYVVREDLVAVE
              :::::::::::::::::::::::::::::::::.::::::::::::::::::::::::::::
SEQ ID NO:181 DFEKGMTAYLDLLLGNCYLMPLNTSIVMPPKNLVELFGKLASGRYLPQTYVVREDLVAVE
                   150       160       170       180       190       200

                       160       170       180       190       200       210
SEQ ID NO:121 EIRDVSNLGIFIYQLCNNRKSFRLRRRDLLLGFNKRAIDKCWKIRHFPNEFIVETKICQE
              ::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::
SEQ ID NO:181 EIRDVSNLGIFIYQLCNNRKSFRLRRRDLLLGFNKRAIDKCWKIRHFPNEFIVETKICQE
                   210       220       230       240       250       260
```


FIGURE 11

```
83.4% identity in 211 aa overlap
                                                      10        20        30
SEQ ID NO:128                                      LWWFWLLWTVLILFSCCCAFRHRRAKLRLQ
                                                   ::::::::::::::::::::::::::::::
SEQ ID NO:182 ELCPGVNTQPYLCETGHCCGETGCCTYYYELWWFWLLWTVLILFSCCCAFRHRRAKLRLQ
                     70        80        90       100       110       120

                     40        50        60        70        80        90
SEQ ID NO:128 QQQRQREINLLAYHGACHGAGPFPTGSLLDLRLLSTFKPPAYEDVVHRPGTPPPPYTVAP
              :::::::::::::::::::::: ::::::::::::.::::::::::.::::::::::.:
SEQ ID NO:182 QQQRQREINLLAYHGACHGAGPVPTGSLLDLRLLSAFKPPAYEDVVHHPGTPPPPYTVGP
                     130       140       150       160       170       180

                     100       110       120       130       140       150
SEQ ID NO:128 GRPLTASSEQTCCSSSSSCPAHFEGTNVEGVSSHQSAPPHQEGEPGAGVTPASTPPSCRY
              : : :.::: : :::.:::.::.:::::::::::::: ::::::: ::..:. .:::::::
SEQ ID NO:182 GYPWTTSSECTRCSSESSCSAHLEGTNVEGVSSQQSALPHQEGEPRAGLSPVHIPPSCRY
                     190       200       210       220       230       240

                     160       170       180       190       200       210
SEQ ID NO:128 RRLTGDSGIELCPCPASGEGEPVKEVRVSATLPDLEDYSPCALPPESVPQIFPMGLSSSE
              ::::::::::::::.:.::::.::::::. :::::.::::::::.::.:. :::::.::
SEQ ID NO:182 RRLTGDSGIELCPCPDSSEGEPLKEARASASQPDLEDHSPCALPPDSVSQVPPMGLASSC
                     250       260       270       280       290       300


SEQ ID NO:128 GDIP
              :
SEQ ID NO:182 GTSHK
```

FIGURE 12